Coefficient of Determination Example

Logistic regression with the natural logarithm base was used to determine the function describing the exponential curve that attempts to fit the density versus line cost data set. To substantiate the accuracy of the resulting data, it appears that the Pearson product-moment correlation coefficient, Pearson's *r*, was used to calculate the coefficient of determination, r^2. The use of Pearson's coefficient to test the strength of correlation between density (*DNS*) and predicted line costs (*LCHat*) simply proves logistic regression methods can fit a curve to a set of data points. This is not an indication of the strength of the correlation between the predicted line costs and the original line costs nor could it result in any meaningful coefficient of determination.

The following example will demonstrate, both visually and mathematically, the fallacy in using Pearson's *r* to substantiate the validity of the predicted line costs (*LCHat*) at a 95% correlation of determination (r^2) (see *NUSF-26.2004.07.08.Erratum to PO No 5 Distribution Model.xls*, Reg_Results worksheet, cell K3). A subset of data from the NUSF-26.2004.07.08.Erratum to PO No 5 Distribution Model was chosen for illustrative purposes. The data records from row 3 through, and including, row 510 on the Reg_Results worksheet were copied to the Data Subset worksheet in the *PBC-1B.xls* file. These records represent the SAM support areas for which 98 percent of the support is received.

Figure 1, below, graphically depicts the relationship between density (*DNS*) and the original line costs (*LC*).

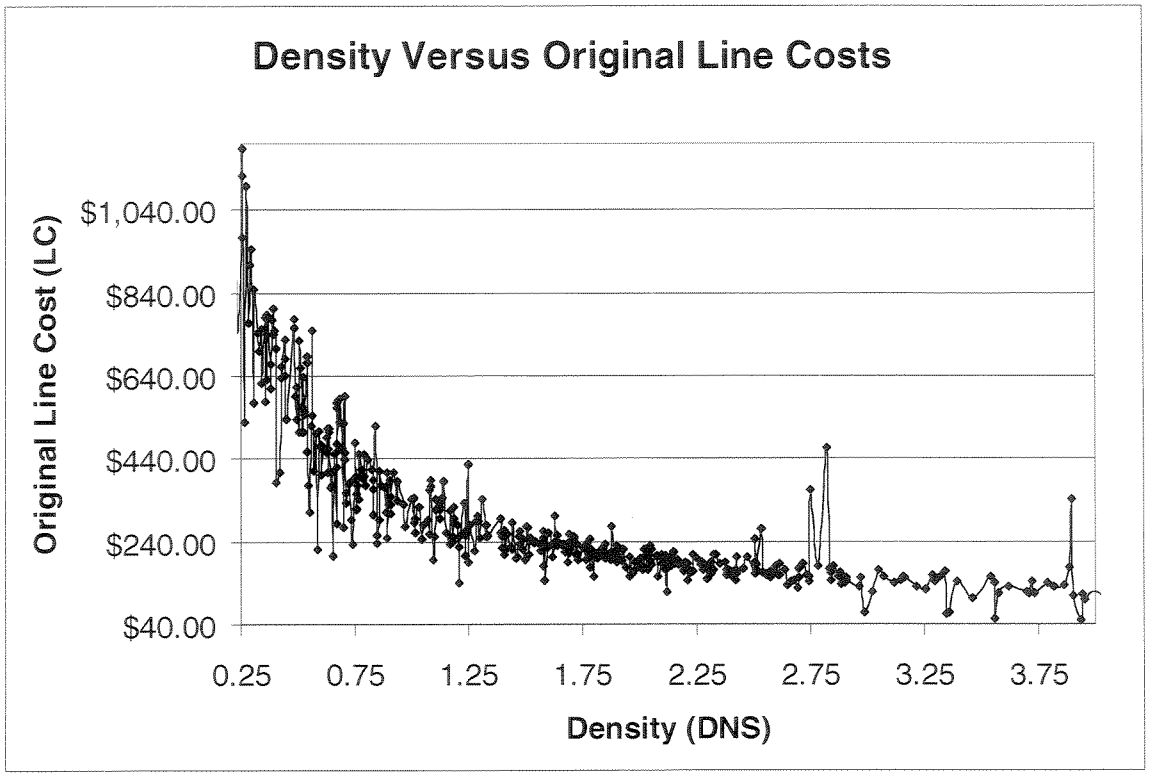

Figure 1: Density Versus Original Line Costs

It is obvious from the scattered points on the graph how difficult it would be to accurately predict the the line cost at a given density. The Pearson product-moment correlation coefficient, *r*, can be used to confirm the prediction difficulty with the data set. Pearson's *r* describes the strength of the relationship between two sets of variables, *X (density)* and *Y (line cost)*.

The equation for Pearson's coefficient of correlation is as follows:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$$

Equation 1: Pearson's Product-Moment Correlation Coefficient

where:

n	is the number of paired observations.
$\sum X$	is the X variable summed.
$\sum Y$	is the Y variable summed.
$(\sum X^2)$	is the X variable squared and the squares summed.

$(\sum X)^2$ is the X variable summed and the sum squared.

$(\sum Y^2)$ is the Y variable squared and the squares summed.

$(\sum Y)^2$ is the Y variable summed and the sum squared.

Pearson's *r* can range from -1.00 to +1.00 where a correlation coefficient of ±1.00 indicates perfect correlation. A correlation around ±0.50 indicates a moderate positive or negative correlation. A correlation close to 0 (either positive or negative) shows that the relationship is quite weak. The strength of the correlation does not depend upon whether the correlation coefficient is positive or negative. A negative correlation coefficient indicates an inverse, or decreasing, relationship.

The calculated Pearson's *r* for the density (*DNS*) and original line cost (*LC*) is -0.63. A Pearson's *r*, or correlation coefficient, of -0.63 indicates moderate, negative correlation between the two sets of variables. Viewing the scattered data points on the graph, one would expect a moderate correlation coefficient for this data set indicating a moderate prediction difficulty.

The calculations are located on the Coefficient Calculations worksheet in the *PBC-1B.xls* file. The calculations were performed two ways with the same results: one using the formula in Equation 1 (cell R6) and a second using Microsoft Excel's® PEARSON function (cell S6).

Next, the same logistic regression form (using the the natural logarithm, e, as the base) used by Staff was employed to determine a function representing a curve for this data. These calculations are located on the Regression Analysis worksheet in the *PBC-1B.xls* file. The curve that results using logistic regression analysis is shown in Figure 2, below.

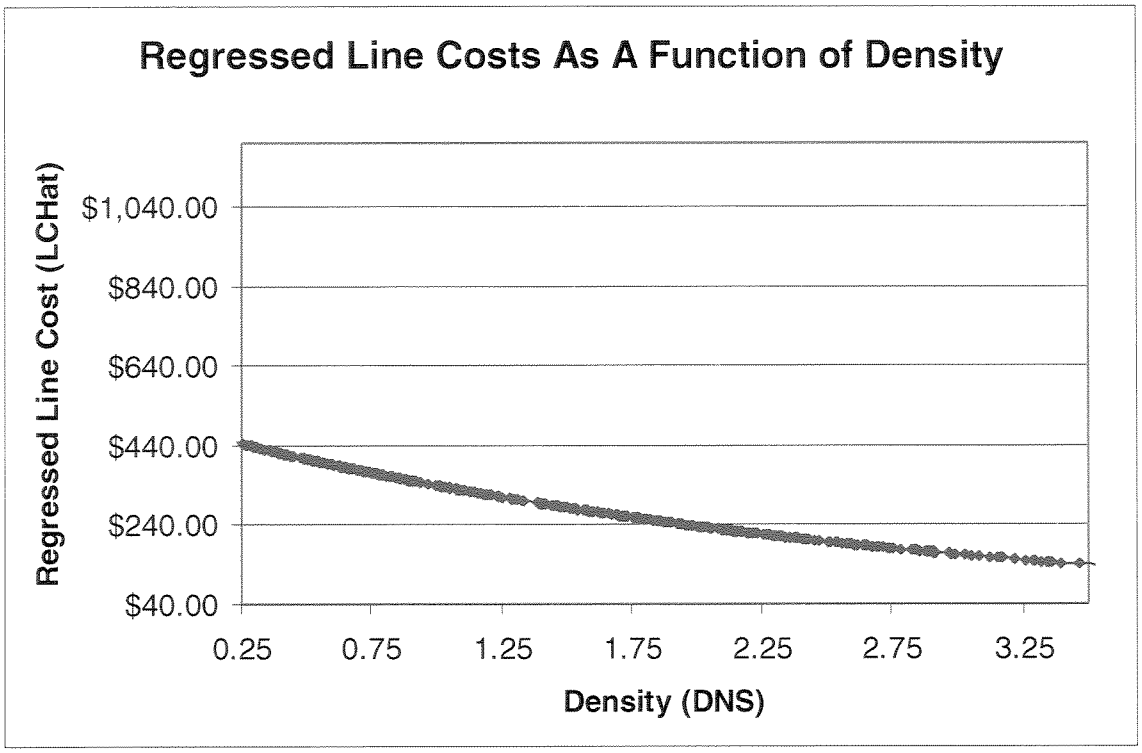

Figure 2: Regressed Line Costs as a Function of Density

The new set of variables consists of the density (DNS) and the predicted (regressed) line cost (LCHat). The graph of the line costs as a function of density is a relatively linear curve representing a much higher correlation between density and the *new* line costs (LCHat). The coefficient of correlation calculated for this new set of variables is -0.95. (See cell T6 in the Coefficient Calculations worksheet in the *PBC-1B.xls* file.) This coefficient of correlation shows a strong relationship as expected, because it merely confirms that regression analysis can produce a function with a very strong, almost perfect, correlation between the two sets of variables (density and the predicted line cost). Just as in the calculation of Pearson's *r* for Figure 1: Density Versus Original Line Costs, the calcuation does not describe the relationship between the original line costs and the predicted line costs, nor does it describe the confidence level of the derived function. In addition, it cannot be used to determine a meaningful coefficient of determination nor the statistical validity of the SAM. The coefficient of determination, r^2, describes the proportion of the total variation in the dependent variable (*LCHat*) that is explained, or accounted for, by the variation in the independent variable (*DNS*).

Up to this point in the example, no function has been derived representing a curve that attempts to predict the original line costs. The predicted line costs as a function of density are plotted in Figure 2 and the graph appears to be relatively linear. Taking the scatter diagram of the original data set in Figure 1 and comparing the diagram of the derived function in Figure 2, it is obvious that in very few cases would the predicted values exactly match the original line costs. (See Figure 3: Compare Original Line Costs and Predicted Line Costs, below.)

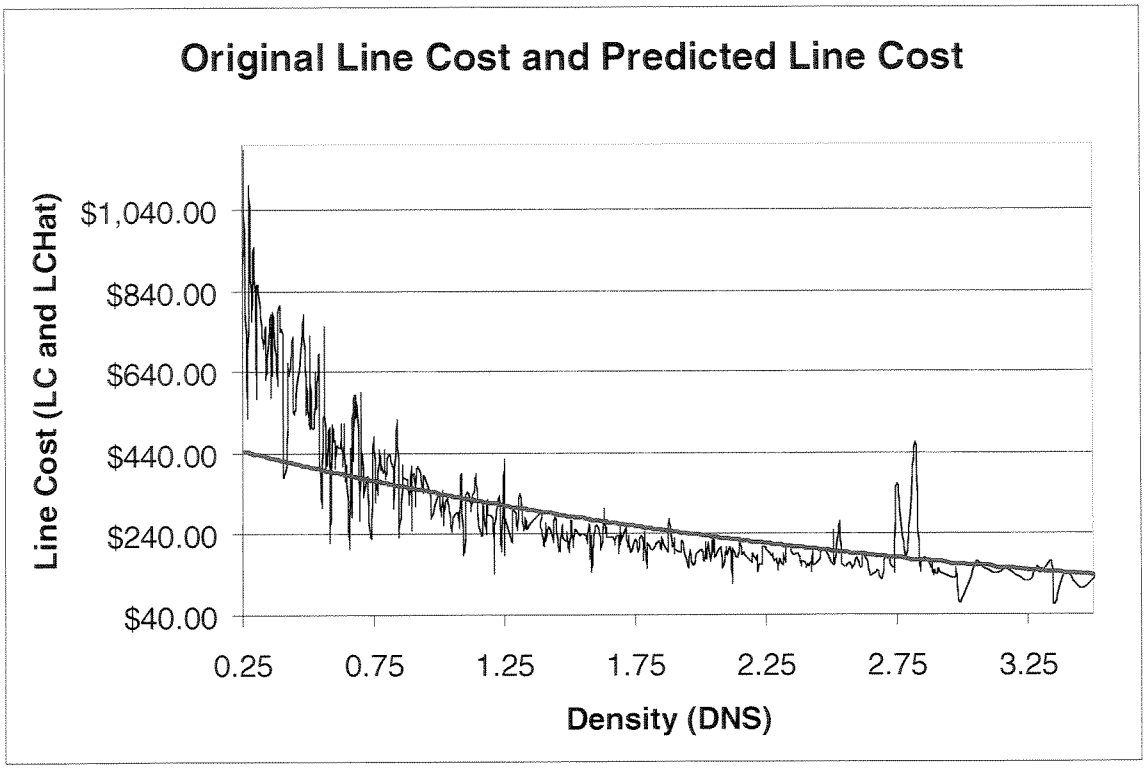

Figure 3: Compare Original Line Costs and Predicted Line Costs

The next step is to test the accuracy of the derived function as compared to the original data set by observing the deviations between the original line costs and the predicted line costs. Regression analysis is a statistical method where the the mean of the random variable (line cost) is predicted conditioned on the independent variable (density), as shown in Figure 1, above. Put another way, this statistical method reduces the data to a straight line, or line of best fit, that attempts to average out the variations in the dependent variable (line cost). Since the new function that is used to predict the random variable (line cost) based on the independent variable (density) is derived by averaging out the variations, it is only logical to test the validity of the resulting function by comparing the variations of the predicted variable relative to the original value with the variations of the mean relative to the original value.

Logically, the *total variation* in line costs is defined as the sum of the *explained variation* and the *unexplained variation* in line costs.

$$\text{Total Variation} = \text{Explained Variation} + \text{Unexplained Variation}$$

Equation 2: Total Variation

The explained variation is defined as the change in the original line costs that is mathematically due to the change in density. The total variation in the predicted line cost data can be derived by subtracting the mean original line cost from the original line cost and summing the squared deviations, $\sum(Y-\bar{Y})^2$. Comparing the mean line cost to the original line cost keeps the sum of the squared prediction errors at a minimum. The total variation in line costs is calculated in column K on the Coefficient Calculations worksheet in the *PBC-1B.xls* file.

To measure the overall error in the predicted line costs ($\hat{Y}$ or LCHat), every deviation from the line should be squared and the squares summed, $\sum(Y-\hat{Y})^2$. This variation cannot be explained by the independent variable and is, therefore, referred to as the *unexplained variation* between the original line costs (Y) and the predicted line costs ($\hat{Y}$). In other words, the unexplained variation is the deviation between the original and the predicted line costs that is not mathematically due to changes in the density. The unexplained variation in line costs is calculated in column N on the Coefficient Calculations worksheet in the *PBC-1B.xls* file.

Subtracting the unexplained variation from both sides of Equation 1, above, gives the equation for the explained variation.

$$\text{Explained Variation} = \text{Total Variation} - \text{Unexplained Variation}$$

Equation 3: Explained Variation

Dividing the explained variation by the total variation gives the percentage of the variation in line costs that is explained, or accounted for, by its linear relationship with density. This is known as the coefficient of determination, r^2.

The coefficient of determination is represented by the following formulas:

$$r^2 = \frac{\text{Total Variation} - \text{Unexplained Variation}}{\text{Total Variation}}$$

$$r^2 = \frac{\sum(Y-\bar{Y})^2 - \sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2}$$

Equation 4: Coefficient of Determination

where:

$\bar{Y}$ is the mean LC.

$\hat{Y}$ is the predicted point LCHat.

Since this formula relates the original line costs to the predicted line costs, this calculation is the appropriate method, both intuitively and mathematically, to determine the confidence level of the regression analysis results. One cannot test the accuracy of a derived function without comparing the predicted values (LCHat) to the original values (LC). The only meaningful way to determine the level of confidence in a derived function is by quantifiying the percentage of expected, or explained, variations to the total variations.

For this example, r^2 is calculated to be 49% (see cell S10 on the Coefficient Calculations worksheet in the *PBC-1B.xls* file). Therefore, only 49% of the total variation in line costs is accounted for by the variation in density – a very low level of confidence. This indicates over 50% of the variation cannot be explained by the variation in density and, thus, represents the overall error percent.

An r^2 of 49% is what would be expected given the deviations between the original line costs and the predicted line costs depicted graphically in Figure 3. An r^2 of 90% (r of 0.95), as calculated for Figure 2, could never be expected to describe the level of confidence for the regressed function as compared to the original function in Figure 3. Both intuitively and mathematically, it is more appropriate to use Equation 4: Coefficient of Determination for determining the level of confidence for a derived function.

This same analysis was applied to the segmented regressional analysis performed by Staff. The coefficients were accurately reproduced within five decimal places using Microsoft Excel's® LOGEST function. The calculations can be seen on the Replicate Reg_Results worksheet in the *PBC-1B.xls* file.